Design of Web Questionnaires

Toepoel, V.; Das, J.W.M.; van Soest, A.H.O.

Publication date:
2005

CentER

Discussion Paper

No. 2005–114

DESIGN OF WEB QUESTIONNAIRES: A TEST FOR NUMBER OF ITEMS PER SCREEN

By Vera Toepoel, Marcel Das, Arthur van Soest

October 2005

ISSN 0924-7815

Tilburg University

Design of Web Questionnaires:

A Test for Number of Items per Screen

Vera Toepoel[*], Marcel Das[*], and Arthur van Soest[**]

Abstract This paper presents results from an experimental manipulation of one versus multiple-items per screen format in a Web survey. The purpose of the experiment was to find out if a questionnaire's format influences how respondents provide answers in online questionnaires and if this is depending on personal characteristics. Four different formats were used, varying the number of items on a screen (1, 4, 10, and 40 items). To test how robust the results were, and to find out whether or not a specific format shows more deviation in answer scores, the experiment was repeated. We found that mean scores, variances and correlations do not differ much in the different formats. In addition, formats show the same deviation of item scores between repeated experiments. In relation to non-response error, we found that the more items appear on a single screen, the higher the number of people with one or more missing values. Placing more items on a single screen a) shortens the duration of the interview, b) negatively influences the respondent's evaluation of the duration of the interview, c) negatively influences the respondent's evaluation of the layout, and d) increases the difficulty in completing the interview. We also found that scrolling negatively influences the evaluation of a questionnaire's layout. Furthermore, the results show that differences between formats are influenced by personal characteristics.

JEL codes: C42, C81, C93

keywords: web survey, questionnaire design, multiple-item-per-screen, scrollable vs. multiple screen, measurement error, non-response error

[*] CentERdata, Tilburg University, postal address: CentERdata, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Corresponding author: Vera Toepoel; e-mail: V.Toepoel@uvt.nl

[**] Tilburg University, Faculty of Economics and Business Administration, Department of Econometrics and Operations Research; RAND, Santa Monica, US.

1. Introduction

Web survey design has had a greater emphasis on programming skills and Web page design than on survey design. In many surveys the response format is chosen for reasons that have little to do with reducing measurement error. For example, matrix questions are frequently used either to save space on the screen or to reduce the number of screens. But does the use of a particular questionnaire format affect the answers to survey questions? Is non-response error enhanced by a questionnaire's format? And how does the respondent evaluate this format?

Respondents appear to see adjacent items (verbally or visually) as more related. This effect of the order of questions indicates respondents' cognitive processing and indirectly illustrates the influence of conversational norms on mental processes (Schwarz and Sudman, 1996).

Literature provides evidence that words and graphical languages combine in ways that influence how people respond to questionnaires (see Couper et al., 2001; Dillman et al., 2003; Lozar Manfreda et al., 2002; Sanchez, 1992); we are as yet some ways from knowing how and why these elements act as they do. Also, when understanding the question-answering process special attention should be given to the questionnaire's format interaction with personal characteristics, such as age groups, education groups and sexes. Research on this matter is still at the frontiers of Web survey methodology.

In this paper the impact of one important Web survey design feature, the use of one versus several items per screen format, will be discussed. First, the questionnaire consisting of forty items is tested on measurement

error (section 4.1). Second, we take a closer look at non-response error in relation to the questionnaire's format (section 4.2). Third, time to complete the questionnaire is taken into account (section 4.3). And fourth, the evaluation of the questionnaire is used in order to get a better understanding of the respondent's experience answering the survey (section 4.4). These four topics were further analyzed to find out in which way questionnaire format interacts with personal characteristics. The study was conducted in the CentERpanel, an online household panel, representative of the Dutch population. To test how robust the results were, and to learn whether or not a specific format shows more deviation in answer scores in repeated experiments, the questionnaire was repeated on the same sample.

2. Background

Survey data are only as meaningful as the answers provided by the survey respondents. Trying to understand how respondents comprehend survey questions leads inevitably to a more basic search for cognitive processes involved in answering questions. Interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, formatting a response, and editing it are the main psychological components of a process that starts with respondent's exposure to a survey question and ends with their report (Sudman et al., 1996). How a respondent processes each of these steps is very context-dependent. The influence of the context in which a question is presented is more pronounced as the question becomes more ambiguous. Even under conditions where respondents can retrieve an opinion on an issue from memory, the opinion

may not exactly match the facet tapped in the question. Similarly, respondents are unlikely to have an appropriate answer to most behavioral questions stored in memory. As a result, most of the given answers in surveys reflect judgments that respondents generate on the spot in the specific context of the specific interview. The research on self-administered surveys suggests that layout and other graphic cues built into the format of questionnaires play an important role in communicating question objectives to respondents (Couper et al., 2001; Dillman et al., 2003; Lozar Manfreda et al., 2002; Sanchez, 1992). Differences in design yield detectable effects.

We distinguish screen-by-screen and scrolling techniques (horizontally/vertically) with respect to navigation in Web survey design. Schonlau et al. (2002) suggest that excessive scrolling can become a burden to respondents and lengthy web pages can give the impression that the survey is too long to complete. On the other hand, scrolling questionnaires can lead to shorter completing times. A disadvantage of the use of screen-by-screen format can be a lack of context. If people are not able to see the entire survey before answering, an order effect can be enhanced.

There are a number of question types used in survey research, such as radio buttons, check boxes, drop boxes, scalar questions and matrix questions. In their search for context effects, Couper et al. (2004) explore three response formats used in Web surveys: a series of radio buttons, a drop box with none of the options initially displayed until the respondent clicks on the box, and a scrollable drop box with some of the options initially visible, requiring the respondent to scroll to see the remainder of the options. They find evidence that visibility may be a more powerful effect than primacy in

Web surveys. They also found support that items earlier in a list are subject to deeper cognitive processing.

Ordinal scale questions require respondents to select a category that best represents where they fit along a continuum from negative to positive. In constructing ordinal scales for self-administered questionnaires, the visual layout of the scale is an important source of information that respondents use when deciding which answer to select. Scalar questions are questions most presented (and most suited) in matrix questions. Presenting a question in a matrix saves space on the screen and reduces the number of screens. Yet, Schonlau et al. (2002) suggest to use matrix questions sparingly. Because matrix questions require a lot of work within a single screen, and make it more difficult to predict how a matrix question will look on a respondent's Web browser.

The grouping of related items on a single screen is likely to lead respondents to view the items as related entities, thus increasing the correlation among them. Adjacent items are more likely to be considered related than items placed at further distance from one another, reflecting a natural assumption that blocks of questions bear on related issues, much as they would during ordinary conversations. Variables that can elicit the application of the conversational norm of nonredundancy include the graphical layout of self-administered questionnaires (Sudman et al., 1996).

Couper et al. (2001) examined in an experiment two types of items on a student population. Each item was measured on a 5-point Likert-type scale. They concluded that correlations are consistently higher among items appearing together on a screen than items separated across several screens.

However, the overall effect is not large, and none of the differences between each pair of correlations reach statistical significance. They also conducted factor analyses of the set of attitude items and found similar factor structures across the one-item-per-screen and several-items-per-screen versions. Thus, a modest support for the grouping hypotheses.

Tourangeau et al. (2004) replicate the above findings on grids versus single items. In their experiment, two of the eight items were ‘reverse worded’. The relation of these two items to overall scores (the part-whole correlations) was weaker when the eight items were presented in a single grid (r=-.331 and r= -.097) than when they were presented in two grids on separate screens (r=-.395 and r=-.151) or on eight separate screens (r=-.427 and r=-.187). Respondents seemed to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully.

Bradlow and Fitzsimons (2001) conducted an experiment using a multi-item scale that consisted of five dimensions and manipulated context effects such as explicit item labeling, item presentation (alone/grouped) and subscale items presented contiguously or not. They suggest that clustering items help maintain the subscale structure. When items are not clustered, items within the same subscales are basically uncorrelated. Also, when items are not labeled or clustered, respondents base their responses on the previous item to a greater degree, regardless of whether the item is intrinsically related.

Non-response error can be divided to unit non-response (a person does not fill in the questionnaire), partial non-response (a person does not finish the questionnaire), and item non-response (a person does not give an answer to a particular item).

Multi-page design can result in early abandoning because responding time may be longer, and more actions need to be taken to answer each question in relation to one-page design. With regard to early abandoning (partial non-response) Lozar Manfreda et al. (2002) find no evidence of differences between one-page and multiple-page design. They do confirm that one-page design results in higher item non-response. Further research is suggested on the optimum number of questions per screen and the impact of number of questions per screen on response.

Item non-response is an indicator of a Web survey's navigability and design. Specifically, item non-response is affected by such factors as inadequate information organization, poor navigational flow, and improperly worded questions. It is possible that the questionnaire's format could lead to increases in non-response error if some people are so uncomfortable with the survey's layout and design that they stop and fail to complete the Web survey (Bowker and Dillman, 2000).

Couper et al. (2001), Lozar Manfreda et al. (2002), and Tourangeau et al. (2004) all found evidence that a multiple-item-per-screen format took less time to complete. They concluded that multiple-item screens may be beneficial, but may also require more careful design given screen limitations and browser variations.

Personalization, precontact, follow-up contact and incentives are the factors most associated with higher response rates in Web studies (Cook et al., 2000; and Kaplowitz et al., 2004). Less attention has been paid to motivating tools inside the survey. Respondent motivation influences the decision to participate in the survey, the involvement answering the survey

and the decision to participate in other surveys. A respondent's motivation might have a direct impact on measurement and non-response error.

Self-administered surveys are written in words and in graphics (visual characteristics). Respondents are guided in their interpretation of both words and graphical language by culture (Dillman et al., 2000). Culture can be defined as the learning that goes on through life leading a person to interpret words and symbols in certain ways. Children grow up with computers nowadays. As a result, it can be expected that children (or young people in general) are influenced differently by questionnaire format and design than older people. Studies show that cognitive abilities (often indicated by respondent education) may affect responses. Furthermore, reduction in cognitive functioning due to the aging process is associated with a decline in the reliability of survey responses (Borgers et al., 2004). Deutskens et al. (2004) and Dillman et al. (2003) conclude future research on Internet-based surveys should be directed at confirming the effects of presentation on different questionnaires and populations with different levels of cognitive ability (preferably an online panel). Bradlow and Fitszsimons (2001) and Dillman et al. (2000) conclude formal experiments need to be conducted in ways that allow various word and visual manipulations to be individually evaluated. It is important that such research is done in order to learn the relative power of context manipulations influencing respondent behavior.

In order to deepen the literature on the subject, this paper investigates the impact of one important Web survey design feature, the use of one versus several items per screen format. This will be discussed in relation to

measurement error, non-response error, time to complete the interview and the evaluation of the questionnaire.

3 Design and Implementation

Researchers always set contexts that influence some aspect of the question answering process, either deliberately but often unintentionally as well. What has been called measurement errors in survey methodology literature can be interpreted as those cases where the context is unthinkingly affected and thus resulting in error. The concept of error is more problematic in attitude measurement. Where reports about behaviors or events can (sometimes) be verified, attitude reports reflect subjective evaluative judgments; there is no objective standard that reflects respondents 'true' attitude.

In order to understand errors due to context effects, our experiment compared several layout options for a questionnaire consisting of 40 items based on a measurement scale developed by Mehrabian and Russell (1974). Details can be found in Appendix A. This scale was constructed to measure arousal. We use this arousal scale because it is an arbitrary, validated scale, which does not vary much in several years. All items were answered on a 5-point Likert scale (1= totally disagree; 2= disagree; 3= disagree nor agree; 4= agree; 5= totally agree). As mentioned earlier, the more ambiguous the question the more likely context effects are to emerge. Preceding questions may influence what comes to mind when respondents anchor a rating scale, thus influencing the response formatting process. In addition, preceding questions may increase or decrease respondents' concerns about self-presentation and social desirability, affecting the editing stage of the question answering process (Sudman et al., 1996).

Literature suggests respondents view items that are clustered, as more related. Therefore, context effects can be enhanced. The aim of the experiment is to investigate the effect of presenting items of a scale on separate screens, several screens or one screen on measurement and response error.

The experiment was conducted in the CentERpanel, an online household panel consisting of more than 2,000 households. This panel is representative for the Dutch population (see Appendix C for more details about the CentERpanel). Because not all people own a computer, CentERdata provides a so-called setop box (and if necessary a television set) to make it possible for them to fill in the questionnaire via the Internet.

Respondents were randomly divided into seven groups.
The first group answered each item on a single screen. The second group answered four items per screen, while the third group answered ten items per screen. The fourth group answered all 40 items on one single screen. Because of the height of the screen, and because of differences in resolutions, people with ten or more items per screen had to scroll in order to fill in all the items.

If the respondent scrolled, he/she was not able to see the header (totally disagree-totally agree) anymore (the header was only programmed at the top of each screen). We also wanted to find out if there exists an extensive difference in the visibility of the header. Thus, 3 more groups were made, another four-, ten- and all-items per screen group, but in these three groups a header at each item was displayed, resulting in four, ten and forty headers per screen, as opposed to one header per screen.

In the analysis we combine the single header and multiple header formats[1], because preliminary analyses show no differences in the visibility of the header. We therefore speak of four different formats (see Table 1 and Appendix B for some screen dumps). The visibility of the header is only mentioned if significant differences between the single header and multiple header formats are found.

[Table 1]

In order to find out how robust our results are, and because we want to know whether or not a format influences a respondent's answer (see Hypothesis 7 below), the experiment was repeated some six months later. In this second experiment we again assigned 4 groups (format 1 to 4), but didn't vary the visibility of the header; we used the single header formats. As one can see in Table 2, 2027 respondents completed both questionnaires.

[Table 2]

Before the experiments were conducted the seven hypotheses listed below were formulated.

Hypothesis 1: The mean overall arousal score differs per format.

[1] As a result, the one-item-per-screen version has half the number of respondents of the other formats.

Because the literature suggests that a questionnaire's format influences the answers by a respondent, the overall score of a scale can be influenced by a questionnaire's format.

Hypothesis 2: The more items appear on a single screen, the higher the average correlation between items.

Existing literature suggest that the more items appear on a single screen, the higher the correlation.

Hypothesis 3: The more items appear on a single screen, the higher the item non-response.

It is expected that the single-item-per-screen group will have the least item non-response, because it is easier to forget to fill in an item if there are forty items on the screen, than if there is one item per screen.

Hypothesis 4: The more items appear on a single screen, the less time it takes to complete the questionnaire.

On each screen there appears a 'next' and a 'back' button. In order to go to the following screen, respondents have to click on the 'next' button. Thus, respondents have to click this button forty times if they have the one-item-per-screen version. But if they fill in the forty-item-per-screen version, they only have to click the button once (though some scrolling is needed).

Hypothesis 5: A format influences a respondent's evaluation of the questionnaire.

At the end people are asked to evaluate the questionnaire. Questions are asked about the duration, the question wording, the easiness and the layout. Because the literature suggests it takes less time to complete the questionnaire when more items appear on a screen, the evaluation of the questionnaire should be better when more items appear on a single screen. On the other hand, the more items within a single screen, the larger the information intake, making a screen more difficult to process. Therefore, different evaluation scores per format are expected, depending on which aspect of the evaluation is being asked.

Hypothesis 6: The effect of different formats on evaluation and response differs for respondents with different personal characteristics.

Some people can handle information easier than others; therefore one can expect different evaluations and responses when looking at age, education, and sex.

Hypothesis 7: The more items appear on a single screen, the more deviation of a respondent's 'true' score (the higher the absolute difference between items in repeated experiments).

To measure if a format influences whether or not a respondent gives his/her true opinion, the same questionnaire at time t+1 is measured. It is expected that people who are uncomfortable with a specific format fail to read the questions properly and therefore give different answers at time t+1 in regard to time t. We expect format 4 (all forty questions on a single screen) to differ most between the repeated measures because this format has the

largest information intake on a single screen. Since the second experiment was held almost six months after the first experiment, it is unlikely that respondents remember exactly the answers given before.

4 Results

In the following sections measurement error (section 4.1), non-response error (section 4.2), time to complete the interview (section 4.3) and the evaluation of the questionnaire (section 4.4) are discussed. Affects of personal characteristics are taken into account into the analyses.

4.1 Measurement error

The counting of all forty items of the scale resulted in an overall score of arousal. We wanted to see whether the mean score of this arousal-scale was the same for all formats. A one-way between-groups analysis of variance was conducted to explore the impact of survey format on the arousal-scale. Analysis on 4 different formats (see Table 1) showed no statistically significant difference in the mean scores. Analysis of variance (counting the variances of all items per respondent) showed no significant difference either. Thus, we did not find evidence for the first hypothesis.

Based on Couper et al. (2001) and Tourangeau et al. (2004) we hypothesized that grouping items on one screen would increase the correlations among them. There are small differences in inter-item correlations when the items were presented one-item-per-screen (Cronbach's alpha of .8801), 4-items-per-screen (alpha of .8849), 10-items-per-screen (alpha of .8871) or all-items-per screen (alpha of .8788). The 10-items-per-screen

format shows the highest inter-item correlation, indicating that items that can be seen in a glance correlate higher.

Some of the items were ‘reverse worded’ (people who agree with the items are arousal averse). The relation of these items to overall scores was weaker when the items were presented on a single screen than when they were presented on separate screens. However, the overall effect is not large, and none of the differences between each pair of correlations reach statistical significance.

The 40 items of the arousal scale were subjected to factor analysis as well. This revealed the presence of eleven components with eigenvalues exceeding 1 in format 1, 3, and 4, but in format 2 (4 items per screen) 10 components with eigenvalues exceeding 1 were revealed. However, the percent of the variance explained by the components did not differ much (components with eigenvalues over 1 explained 58% of the variance in the one-item-per-screen format, 54% in the 4-items-per-screen format, 57% in the 10-items-per-screen format and 55% in the all-items-per-screen format). Thus no support was found for the grouping hypothesis.

We wanted to find out if including variables for format provides better predictions of the total variance in item scores than excluding these variables. In the regression analysis we took format 1 as reference level. We explained the variance of all items (per respondent) by a linear regression on sex, education and age. We examined the First-Order Model (excluding the dummy variables for formats 2 to 4) and the Second-Order Model (including these variables). We concluded that the second-order terms do contribute to the model ($p=0.00$). Including variables for format provides better predictions.

In addition, we wanted to check if the inclusion of interaction terms of sex, education, and age with the different formats provides better predictions of the total variance in item scores than excluding these variables. To answer this, we examined the First-Order Model (regression on sex, education, age, and dummy variables for formats 2 to 4) and the Second-Order Model (adding the interaction terms). Again, we concluded the second-order terms do contribute to the model (p=0.00). This full model, presented in Table 3, shows a significant first-order and a second-order effect. The higher the age of the respondent, the higher the variance in item scores. Furthermore, women seem to have a higher variance in answering format 2 than format 1 than men.

[Table 3]

To measure if a format influences whether or not a respondent gives his/her true opinion, we fielded the same questionnaire at time t+1. Our hypothesis was that a format influences a respondent answer (an uncomfortable format shows more deviation from the true score).[2]

First, to detect a learning effect (because the respondent answered the same questionnaire half a year after the first experiment), a one-way repeated measure ANOVA was conducted to compare scores on the arousal scale at Time 1 and Time 2. The means and standard deviations are presented in Table 4. No significant effect for time was found. Therefore, we concluded that

[2] In this analysis only the 2027 respondents who answered the questionnaire at Time 1 and Time 2 were taken into account.

if differences between experiments are found, they can be attributed to changes in format.

[Table 4]

In order to find out if a particular format causes more deviations from the true score than another format, we computed a new variable, the "deviation score," indicating the absolute deviation of an answer to Item Y at Time 2 in relation to Time 1.

At Times 1 and 2 each respondent was randomly assigned one of the four formats. Hence the number of format combinations is, $4 \times 4 = 16$. We call respondents with the same format combination a group. A one-way between-groups analysis of variance was conducted to explore the impact of format on the deviation score. No statistically significant difference in the mean deviation in scores for the 16 different groups was found. Therefore, there is no support for the seventh hypothesis. The deviation of a respondent's 'true' score does not differ per format.

4.2 Non-response

Non-response can be divided into unit non-response, partial non-response, and item non-response. Because respondents were randomly assigned to a particular question format after opening the questionnaire, unit non-response[3] is not taken into account. And because only a small number of respondents[4] did not finish the questionnaire, no meaningful conclusions can be made on partial non-response.

No control for item non-response was used in the different formats, so people could proceed in the questionnaire without filling in answers. The difference in item non-response can therefore be attributed to the difference in the layout.

Two different analyses in relation to non-response error were performed. First, the missing items per respondent were counted. Second, a dummy variable for item non-response with 0= respondent had no item missing and 1= respondent had one or more missing items was made. Where the first analysis reveals *how many* missing values exist in the dataset, the second analysis shows *the number of respondents* who had one or more missing items. To see how robust the results are, item non-response with the repeated experiments was analyzed. Table 5 shows the results for the different analyses.

We found in our first experiment that the more items appear on a single screen, the higher the number of people with one or more missing values. There exists a linear relationship between the number of items per screen and

[3] Unit non-response is 31% at Time 1 and 35% at Time 2.
[4] 16 respondents (N=2565) did not finish the questionnaire at Time 1 and 31 respondents (N=2350) did not finish the questionnaire at Time 2.

item non-response. The all-items-per-screen version contains the most item non-response in relation to the one-item-per-screen version with the least item non-response. This is what we expected because the amount of information on a single screen is much higher in the all-items-per-screen version, so it is easier to forget to fill in an answer than if there is only one item to fill in on a screen. Post-hoc comparisons using the adjusted Tukey HSD[5] test (Games-Howell) indicated that the number of item non-response for format 4 (all-items-per-screen) was significantly different from format 1 (1-item-per-screen).

When we look at the dummy variable for non-response the post-hoc comparisons show statistically significant differences for all formats. Again, the differences between the different formats are large. So, not only the amount of missing items increases as more items appear on a single screen, also the number of respondents with one or more missing items increases.

In our repeated experiment similar results were found.

[Table 5]

A linear regression (with format 1 as reference level) of non-response on sex, education, age and interaction terms of these variables with dummy variables for format 2 to 4 revealed no first order (personal characteristics) or second order (interaction terms) effects.

[5] Games and Howell's modification of Tukey's HSD is a modified HSD test that is appropriate when the homogeneity of variances assumption is violated.

4.3 Time to complete the interview

When items are presented on a grid, a respondent has to make fewer physical actions (clicking on the mouse) then when items are presented separately. Therefore, our hypothesis was the more items appear on a single screen, the less time it takes to complete the questionnaire. To test this hypothesis we did a one-way between-groups analysis of variance. Again, we repeated our analysis for Time 2.

In our first experiment we found a significant difference ($p<0.05$) in mean duration of the interview per format after deleting outliers. As one can see in Table 6, the 10-items-per-screen format took the least time to complete at Time 1, followed by the all-items-per-screen format, the 4-items-per-screen format and the 1-item-per-screen format. Post-hoc comparisons using the adjusted Tukey HSD (Games-Howell) showed significant group differences between format 1 (one-item-per-screen) and format 3 (10-items-per-screen), indicating that the survey format has an effect on the duration of the interview.

[Table 6]

The hypothesis 'the more items appear on a single screen, the less time it takes to complete the questionnaire' is true in our first experiment, but does not hold for more than 10 items[6]. Because format 3 seems to have

[6] One can interpret this result in the way that the extensive scrolling in the all-items-per-screen version (to see the header at the top of the screen and then go back in order to fill in the answer) is causing the increase of duration in filling in the interview after 10 items. Further analyses of the visibility of headers (one header per screen vs. one header per item) showed

shorter completion times than format 4, and the group differences between format 3 and format 4 are not statistically significant, more evidence is needed for hypothesis 4.

Our repeated experiment shows more indisputable evidence for our hypothesis. A linear relationship between the number of items appearing on a single screen and the duration of the interview is found. The more items appear on a single screen, the shorter the duration of the interview.

In order to find out if there exist an interaction effect of format and personal characteristics in relation to the duration of the interview, we conducted a regression analysis. No second order effects were found. We do find a significant first order effect for age. The higher is the age of the respondent, the longer the duration of the interview.

4.4 Evaluation of the questionnaire

At the end of the questionnaire, people answered some evaluation questions.

1. How do you evaluate the duration of the interview?
2. How clear was the question wording?
3. How easy were the questions to fill in?
4. How was the layout?

---

that in versions where scrolling is necessary, adding headers per item decreases the time to complete the interview. Therefore, scrolling upwards in order to see the header at the top of the screen can be a cause for the increase in time to complete the interview. However, the differences do not reach statistical significance.

5. What is your overall opinion about this interview?

These questions were asked on a ten-point scale ranging from 1 (‘very bad’) to 10 (‘very good’).

Because we wanted to see how robust our results are, we analyzed these evaluation questions with the repeated experiments (Time 1 and Time 2). A closer look at question 2 (How clear was the question wording?) showed no significant differences between the 4 different formats. In addition, the overall opinion about the interview (question 5) showed no significant difference for all formats either. We found differences between formats for the evaluation of the duration of the interview (4.4.2), the easiness to fill in the answers (4.4.3) and the layout (4.4.4). First we take a closer look at the overall evaluation of the questionnaire.

4.4.1 Overall evaluation

To find out which evaluation question has the highest influence on the overall evaluation of the interview, a regression analysis was conducted. Table 7 shows the coefficients of all evaluation questions on the overall opinion about the interview (question 5). The effect of question 1 (the duration of the interview) is the highest.

[Table 7]

4.4.2 Duration

We only find different evaluation scores for the duration of the interview in our second experiment. While log-files show significant differences in

duration of the interview of the different formats in both experiments, respondents do not evaluate the duration significantly different at Time 1. In our second experiment respondents do evaluate the duration of the interview different for all formats (see Table 8).

[Table 8]

The 4-items-per-screen version (format 2) is evaluated best, followed by the 10-item-per-screen format (format 3), the all-items-per-screen version (format 4) and the one-item-per-screen version (format 1) respectively. Post-hoc comparisons show significant group differences for formats 1, 2 and 4. Although log files show that the all-items-per-screen version took the least time to finish at Time 2, the evaluation of the duration of the interview in this format is worse than the other multiple-items-per-screen versions. This indicates a difference in real and experienced duration of the interview.

4.4.3 Easiness

The question about the easiness to fill in the questions shows different results for Time 1 and Time 2 as well. Where we did not find significant differences in evaluation scores for this question at Time 1, we do find differences for Time 2. There was a statistically significant difference at the $p<0.05$ level between the different formats as one can see in Table 9. Post-hoc comparisons using the adjusted Tukey HSD (Games and Howell) test indicated that the mean score for formats 1, 3 and 4 differ significantly. The more items appear on a single screen, the more difficulty people experienced in answering the questionnaire.

[Table 9]

4.4.4 Layout

The evaluation of layout differs significantly for both experiments (Time 1 and Time 2). There was a statistically significant difference at the $p<0.05$ level in the evaluation of layout between the different formats as one can see in Table 10.[7] Post-hoc comparisons using the adjusted Tukey HSD (Games and Howell) test indicated that the mean score for all groups differ significantly. The more items appear on a single screen, the lower the evaluation of the layout.

[Table 10]

We can conclude that respondents consider the layout of the all-items-per-screen version significantly worse than all other formats, but a mean score of 7 out of 10 is not a reason to stop the use of a scrollable format. We can say, however, that the division of text on a screen is evaluated best in the 1-item-per-screen format followed by the 4-items-per-screen format.

Including the number of headers in the analysis, we find another significant difference. For the 10-items-per-screen version with 10 headers, and the all-items-per-screen version with 1 and 40 headers, respondents had to scroll in order to see the remainder of the screen. A one-way between-groups analysis of variance shows that the 10-items-per-screen version with 10 headers can be grouped with the all-items-per-screen versions. These 3

[7] Because the repeated measure shows similar results as the first measure, we only present the outcome of the one-way between-groups analysis of variance for Time 1.

formats are evaluated significantly worse than the other formats. While scrolling for the 10 items per screen version with 10 headers is absolutely necessary (as opposed to the 10 items per screen version with a single header, where scrolling depends on the screen resolution), one can conclude that whether or not a respondent has to scroll in order to see all items on a screen, affects the evaluation of the screen's layout.

Because a panel is used, the evaluation of layout for a format per person can be compared.[8] With a mean score of 7.42 at Time 1 and a mean score of 7.44 at Time 2, the mean score of the evaluation of layout is the same both times. Respondents, who answered a format with 10 or more items on a single screen, evaluate the layout worse than this mean score.

A one-way repeated measures ANOVA was conducted to compare the evaluation of layout at Time 1 and at Time 2. There was a significant difference (p=0.00), which means that on average people evaluated their format at Time 2 different than their format at Time 1.

We found evidence that respondents evaluate the format worse when more items appear on a single screen. As one can see in Table 11, respondents who had a format at Time 2 containing *more* items on a single screen than the format they answered at Time 1 evaluated the format at Time 2 *worse* than the format they answered at Time 1.

Respondents who had a format at Time 2 containing *less* items on a single screen than the format they answered at Time 1 evaluated the format at Time 2 *better* than the format they answered at Time 1.

[8] Respondents were randomly assigned to the different formats at Time 1 and Time 2. Hence, respondents could have a different or the same format at Time 1 and Time 2.

Respondents who answered the same format at Time 1 and Time 2 show similar evaluation scores for both experiments.

[Table 11]

We wanted to find out if including variables for format provides better predictions of evaluation of layout than excluding these variables.
In the regression analysis we again took format 1 as reference level. We explained the evaluation of layout by a linear regression on sex, education and age. The First-Order Model (excluding the dummy variables for formats 2 to 4) and the Second-Order Model (including these variables) were examined. It was concluded that the second-order terms do contribute to the model ($p=0.00$). Including variables for format provides better predictions for the evaluation of the questionnaire's layout.

In addition, we wanted to check if the inclusion of interaction terms of sex, education, and age with the different formats provide better predictions of the evaluation of layout than excluding these variables. To answer this question, the First-Order Model (regression on sex, education, age, and dummy variables for formats 2 to 4) and the Second-Order Model (adding the interaction terms) were examined. Again, it could be concluded that the second-order terms do contribute to the model ($p=0.00$). Table 12 shows a significant effect of age and format 4 in the full model. The higher the age of the respondent, the better the evaluation of layout in format 4 in relations to format 1. It seems that the elderly do not have a higher preference for a simple style layout: the all items per screen version is evaluated better the

older respondents are. Further, the first-order effect AGE has a significant effect on the evaluation of layout; the higher the age of the respondent, the better the evaluation. The same results were found at Time 2, indicating that our results are robust.

[Table 12]

5 Discussion and Conclusions

In this paper we have investigated, experimentally, how a specific aspect of the visual design and layout of questions influences how respondents process and provide answers in Web surveys and if this depends on personal characteristics. Furthermore, the evaluation of specific formats by the same person is examined, in order to find out which format is evaluated best.

Our focus is on understanding a specific aspect, the use of one versus multiple-items per screen format, to find out if layout changes a respondent's answer. This can be kept in mind when choosing a format (or advice a designer) so measurement errors and non-response errors can be minimized, and financial gains and motivational tools can be enhanced. After all, a researcher's goal is not only to get answers, but also to motivate a respondent to participate actively in a survey.

Our results in relation to measurement error are in the same line with the findings of Couper et al. (2001) and Tourangeau et al. (2004). Mean scores of the arousal scale and variances do not differ in the different formats. The responses to the forty questions were somewhat evenly intercorrelated

when items were presented one, 4, 10 or all-items-per-screen. Even items that were ‘reverse worded’ showed no statistical significant differences among items appearing together on a screen than items presented on several screens. We also found that different formats show the same deviation in item scores between repeated experiments. And while one could expect differences for old versus young people, people with low versus high education and men versus women we only found that women have a higher variance in answering the 4-items-per-screen version in relation to the one-item-per-screen version. The inclusion of interaction terms of sex, education, and age with the different formats do provide better predictions of the total variance in item scores than excluding these variables.

In relation to non-response error, we confirmed the results of Lozar Manfreda et al. (2002). We found that the more items appear on a single screen, the higher the number of missing items. We did not find interaction effects of format and personal characteristics in relation to non-response error.

We found the hypothesis ‘the more items appear on a single screen, the less time it takes to complete the questionnaire’ to be confirmed, but in our first experiment the time to complete the interview gets longer after 10 items. Our second experiment shows no optimum number for items appearing together on a screen (we find a linear result where the all-items-per-screen version had the shortest completion time). Older respondents took more time to complete the interview than younger respondents. We did not find interaction effects of format and personal characteristics.

Keeping in mind that a respondent's motivation influences the decision to participate in the survey, the involvement answering the survey and the decision to participate in other surveys, we saw the more items appear on a single screen, the lower the evaluation of the layout. Including the number of headers in the analysis (one header per item vs. one header per page) we found that scrolling negatively influences the evaluation of a questionnaire's layout. A remarkable result is the fact that the all-items-per-screen format is evaluated better the older a respondent is. Future research on the elderly population could further deepen our understanding in which way reduction in cognitive functioning is associated with a decline in the reliability of survey responses (Borgers et al., 2004) and a preference for layout. We found significant effects in the evaluation of the duration of the interview and the easiness to fill in the answers in our second experiment. It shows that a) respondents evaluate the duration of the questionnaire worse the more items appear on a screen and b) respondents find the questions more difficult to fill in the more items appear on a screen.

Thus differences in item non-response, time to complete the interview, and evaluation scores between the four formats are found, but little evidence was found on a questionnaire's format influencing measurement error. The effect of this specific visual manipulation on the quality of the data is small, but big in regard to item non-response and evaluation of the questionnaire. There seems to be a difference in facts and feelings. While log files show the more items appear on a single screen, the less time it takes to complete the questionnaire, respondents think the duration of the interview is longer the

more items appear on a single screen. With regard to populations with different levels of cognitive ability, some effects were found.

This experiment is only a small step further in a greater understanding of how design features influence respondents. The effects of such features on respondents are clearly worth additional research. Our findings also hint at the possible design trade-offs that could be explored. Optimizing the format-respondent relation can keep the respondent motivated, shorten the interview time and reduce measurement error and item non-response.

6 References

Borgers, Natascha, Joop Hox, and Dirk Sikkel. 2004. “Response Effects in Surveys on Children and Adolecents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point.” *Quality & Quantity* 38: 17-33.

Bowker, Dennis, and Don A. Dillman. 2000. *An experimental evaluation of left and right oriented screens for Web questionnaires.* Paper presented at the 55th annual conference of American Association for Public Opinion Research. Portland, Oregon, May 18-21, 2000. http://survey.sesrc.wsu.edu/dillman/papers/AAPORpaper00.pdf

Bradlow, Eric T., and Gavan J. Fitzsimons. 2001. “Subscale distance and item clustering effects in self-administered surveys: A new metric.” *Journal of Marketing Research* 38: 254-262.

Cook, Coleen, Fred Heath, and Russell Thompson. 2000. “A meta-analysis of response rates in web- or internet-based surveys.” *Educational and Psychological Measurement* 60: 821-836.

Couper, Mick P., Roger Tourangeau, and Frederick G. Conrad. 2004. “What They See Is What We Get. Response Options for Web Surveys.” *Social Science Computer Review* 22: 111-127.

Couper, Mick P. , Michael W. Traugott, and Mark J. Lamias. 2001. “Web Survey Design and Administration.” *Public Opinion Quarterly* 65: 230-253.

Deutschkens, Elisabeth, Ko De Ruyter, Martin Wetzels, and Paul Oosterveld. 2004. “Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study.” *Marketing Letters* 15: 21-36.

Dillman, Don A., Scot Caldwell, and Mary Gansemer. 2000. *Visual Design Effects on Item Nonresponse to a Question About Work Satisfaction That Precedes the Q-12 Agree-Disagree Items.* Paper supported by the Gallup Organization and Washington State University http://survey.sesrc.wsu.edu/dillman/papers.htm.

Dillman, Don. A., Jolene D. Smyth, Leah M. Christian, and Michael J. Stern. 2003. *Multiple Answer Questions in Self-Administered Surveys: The Use of Check-All-That-Apply and Forced-Choice Question Formats.* Draft Paper, Washington State University http://survey.sesrc.wsu.edu/dillman/papers.htm.

Kaplowitz, Michael D., Timothy D. Hadlock, and Ralph Levine. 2004. "A comparison of web and mail survey response rates." *Public Opinion Quarterly* 68: 94-101.

Lozar Manfreda, Katja, Zenel Batagelj, and Vasja Vehovar. 2002. "Design of Web survey questionnaires: Three basic experiments". *Journal of Computer-Mediated Communication* 7, 3 http://www.ascusc.org/jcmc/vol7/issue3/vehovar.html

Mehrabian, Albert, and James Russell. 1974. *An approach to environmental psychology*. Cambridge, MIT Press.

Sanchez, Maria E.. 1992. "Effects of Questionnaire Design on the Quality of Survey Data." *Public Opinion Quarterly* 56:206-217.

Schonlau, Matthias, Ronald D. Fricker, and Marc N. Elliott. 2002. *Conducting Research Surveys via E-mail and the Web*. RAND, Santa Monica.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking About Answers*. Jossey-Bass Publishers, San Francisco.

Schwarz, Norbert, and Seymour Sudman. 1996. *Answering Questions*. Jossey-Bass Publishers, San Francisco.

Tourangeau, Roger & Mick P. Couper, and Frederick Conrad. 2004. "Spacing, position, and order. Interpretive heuristics for visual features of survey questions." *Public Opinion Quarterly* 68:368-393.

Appendix A

The Arousal Seeking Tendency (Mehrabian and Russel, 1974) is a scale designed to measure involvement. The list of 40 items is presented below.

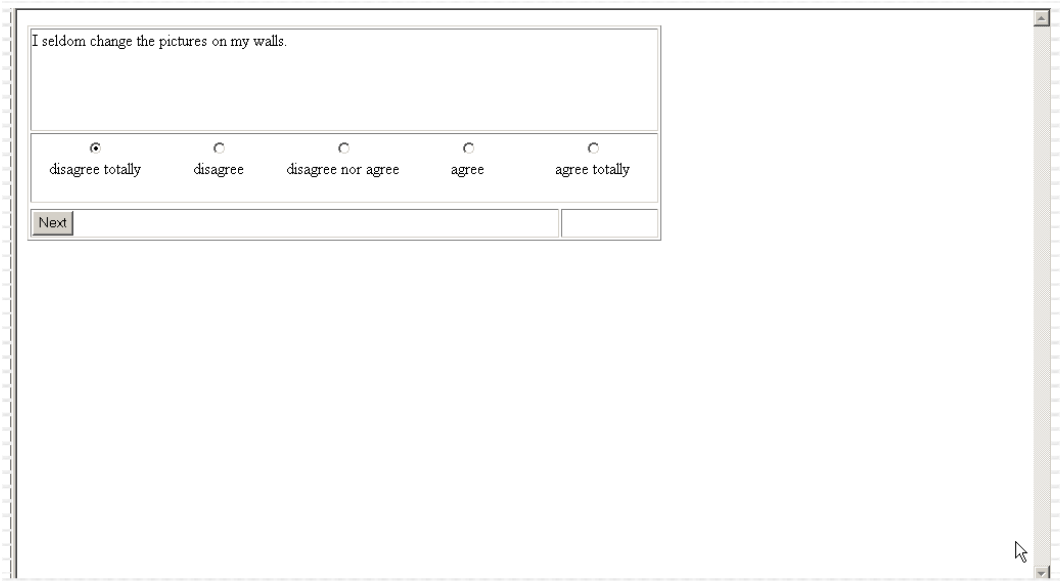
1. I seldom change the pictures on my walls.*
2. I am not interested in poetry.*
3. It is unpleasant seeing people in strange weird clothes.*
4. I am continually seeking new ideas and experiences.
5. I much prefer familiar people and places.*
6. When things get boring, I like to find some new and unfamiliar experience.
7. I like to touch and feel a sculpture.
8. I don't enjoy doing daring foolhardy things just for fun.*
9. I prefer a routine way of life to an unpredictable one full of change.*
10. People view me as quite an unpredictable person.
11. I like to run through heaps of fallen leaves.
12. I sometimes like to do things that are a little frightening.
13. I prefer friends who are reliable and predictable to those who are excitingly unpredictable.*
14. I prefer an unpredictable life full of change to a more routine one.
15. I wouldn't like to try the new group therapy techniques involving strange body sensations.*
16. Sometimes I really stir up excitement.
17. I never notice textures.
18. I like surprises.
19. My ideal home would be peaceful and quiet.*
20. I eat the same kind of food most of the time.*
21. As a child, I often imagined leaving home just to explore the world.
22. I like to experience novelty and change in my daily routine.
23. Shops with thousands of exotic herbs and fragrances fascinate me.
24. Designs and patterns should be bold and exciting.
25. I feel best when I am safe and secure.*
26. I would like the job of a foreign correspondent of a newspaper.
27. I don't pay much attention to my surroundings.*
28. I don't like the feeling of wind in my hair.*
29. I like to go somewhere different nearly every day.
30. I seldom change the décor and furniture arrangement at my place.*
31. I am interested in new and varied interpretations of different art forms.
32. I wouldn't enjoy dangerous sports such as mountain climbing, airplane flying, or sky diving.*
33. I don't like to have to have lots of activity around me.*
34. I am interested only in what I need to know.*
35. I like meeting people who give me new ideas.
36. I would be content to live in the same house the rest of my life.*
37. I like continually changing activities.
38. I like a job that offers change, variety, and travel even if it involves some danger.
39. I avoid busy, noisy places.*
40. I like to look at pictures that are puzzling in some way.

* Reverse-scored

Appendix B

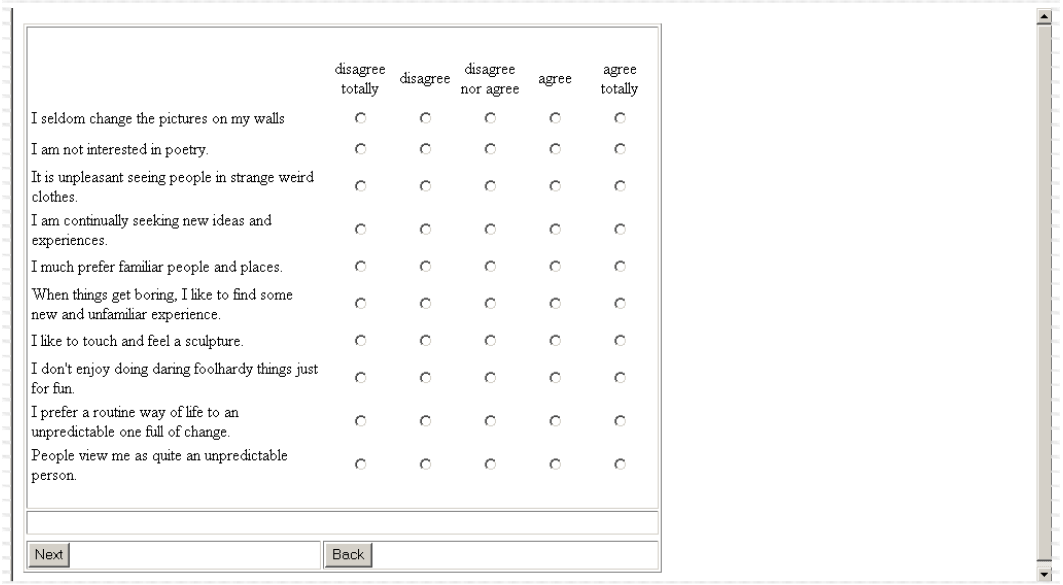
This appendix presents screen dumps for the different formats as were used in the experiment.

Format 1: 1 item per screen



Format 2: 4 items per screen

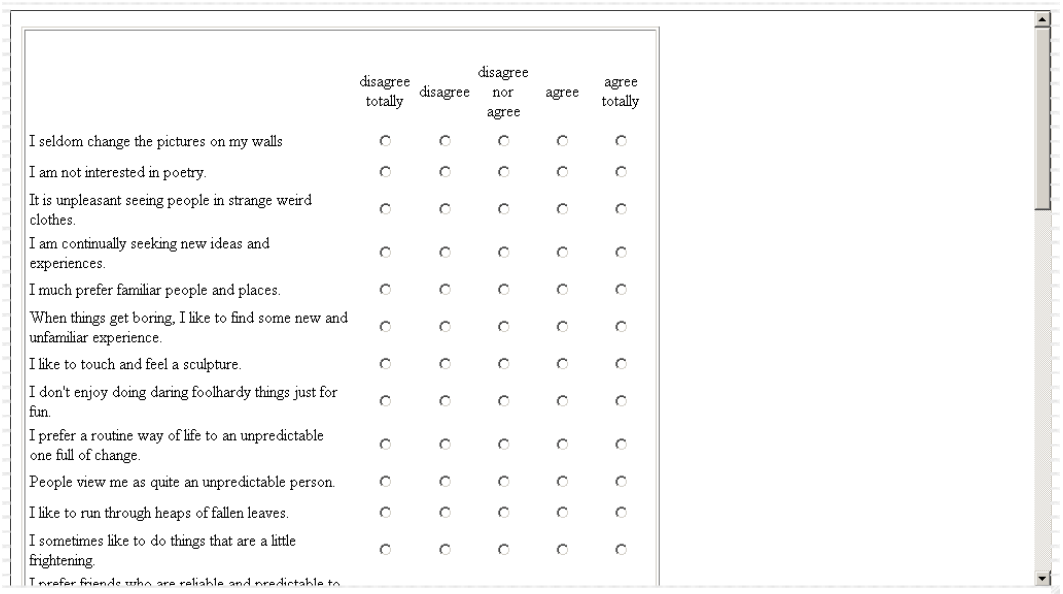
Format 3: 10 items per screen



	disagree totally	disagree	disagree nor agree	agree	agree totally
I seldom change the pictures on my walls	○	○	○	○	○
I am not interested in poetry.	○	○	○	○	○
It is unpleasant seeing people in strange weird clothes.	○	○	○	○	○
I am continually seeking new ideas and experiences.	○	○	○	○	○
I much prefer familiar people and places.	○	○	○	○	○
When things get boring, I like to find some new and unfamiliar experience.	○	○	○	○	○
I like to touch and feel a sculpture.	○	○	○	○	○
I don't enjoy doing daring foolhardy things just for fun.	○	○	○	○	○
I prefer a routine way of life to an unpredictable one full of change.	○	○	○	○	○
People view me as quite an unpredictable person.	○	○	○	○	○

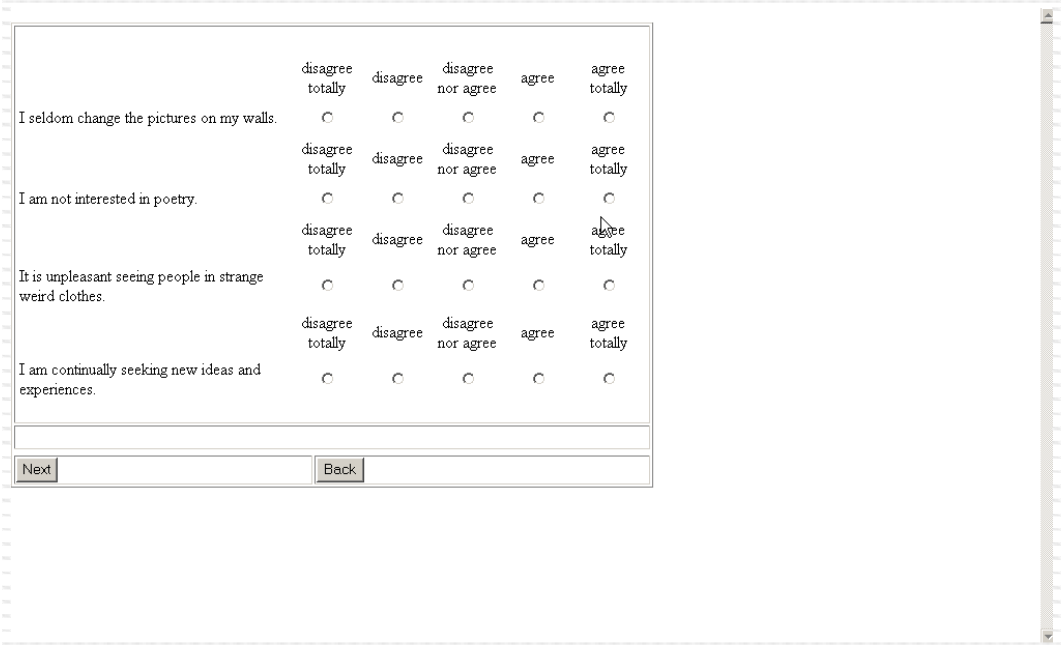
Next Back

Format 4: all items per screen (with the scroll bar on the right-hand side)



	disagree totally	disagree	disagree nor agree	agree	agree totally
I seldom change the pictures on my walls	○	○	○	○	○
I am not interested in poetry.	○	○	○	○	○
It is unpleasant seeing people in strange weird clothes.	○	○	○	○	○
I am continually seeking new ideas and experiences.	○	○	○	○	○
I much prefer familiar people and places.	○	○	○	○	○
When things get boring, I like to find some new and unfamiliar experience.	○	○	○	○	○
I like to touch and feel a sculpture.	○	○	○	○	○
I don't enjoy doing daring foolhardy things just for fun.	○	○	○	○	○
I prefer a routine way of life to an unpredictable one full of change.	○	○	○	○	○
People view me as quite an unpredictable person.	○	○	○	○	○
I like to run through heaps of fallen leaves.	○	○	○	○	○
I sometimes like to do things that are a little frightening.	○	○	○	○	○
I prefer friends who are reliable and predictable to					

Multiple headers per screen (for format 2: 4 items per screen)



NOTE: how the questionnaire exactly appears on screen dependents on the screen height and screen resolution.

Appendix C

This Appendix presents the selection procedure of panel members. The CentERpanel consists of over 2000 households in the Netherlands, the members of which fill in a questionnaire at their home computers every week. The CentERpanel is representative of the Dutch population.

The recruitment of new panel members consists of several stages. In the first stage, a random sample of candidates is interviewed by telephone. In the first telephone interview a number of questions are asked about the demographic characteristics of the household. The interview is concluded with the question whether the person would like to participate in survey research projects. If so, the household is included in a database of potential panel members.

If a household drops out of the panel, a new household is selected from the database of potential panel members. This is done on the basis of demographic characteristics (such that the panel will remain representative of the Dutch population). The selected household is asked whether the members of the household would like to become panel members, and if so, a number of additional questions are asked

Although the CentERpanel is an Internet-based panel, there is no need to have a personal computer with an Internet connection. Those households who don’t have access to Internet, are provided with a so called settop box, with which a connection can be established via a telephone line and a television set. If the household doesn’t have a television, CentERdata provides one also.

Table 1. Different formats in the first experiment

Format		N		*Consists of:*	*N*
1	1 item per screen	352		*1 item per screen*	*352*
2	4 items per screen	727	*a*	*4 items per screen with single header*	*353*
			b	*4 items per screen with multiple headers*	*374*
3	10 items per screen	768	*a*	*10 items per screen with single header*	*370*
			b	*10 items per screen with multiple headers*	*398*
4	40 items per screen	718	*a*	*40 items per screen with single header*	*359*
			b	*40 items per screen with multiple headers*	*359*
	Total	2565			*2565*

Table 2. Response repeated experiments

	Date of fieldwork	N (respons %)	Respondents who filled in both questionnaires
Time 1	December 2004	2565 (69%[9])	2027
Time 2	June 2005	2350 (65%)	2027

[9] Response Rate 1 defined in the Standard Definitions of AAPOR (www.aapor.org)

Table 3. Total variance by linear regression on sex, education, age and interaction terms

N=2565	Parameter	Std. Error	t
Constant**	.902	.131	6.891
Format2	-.316	.167	-1.893
Format3	-.059	.160	-.369
Format4	.090	.163	.552
Sex	.002	.052	.039
Education	.012	.017	2.672
Age*	.004	.002	.685
Format2-Sex*	.169	.064	2.650
Format3-Sex	.093	.063	1.473
Format4-Sex	.027	.064	.415
Format2-Educ	.010	.021	.469
Format3-Educ	-.012	.021	-.559
Format4-Educ	-.035	.021	-1.667
Format2-Age	.002	.002	.836
Format3-Age	.000	.002	-.152
Format4-Age	.001	.002	.518

*p<.01

**p<0.001

Note:

sex:

1=man, 2=woman

education:

1= primary education

2= lower secondary education

3= higher secondary education

4= intermediate vocational training

5= higher vocational training

6= university

Table 4. One-way repeated measures ANOVA, mean and standard deviation for 2 experiments (N=2027)

	Mean	Std.
Time 1	117.95	17.123
Time 2	117.65	17.530

Table 5. Number of item nonresponse (A) and respondent item non-response (B), mean and standard deviations at Time 1.

A) Number of item-nonresponse				
		Mean	N	Std.
Format	1*	0.14	352	0.51
	2*	0.17	727	0.54
	3	0.22	768	0.80
	4*	0.25	718	0.60
	Total	0.20	2565	0.64
ANOVA	F=2.92	p=.03		
B) Respondent item non-response 0= no item missing 1= one or more items missing				
		Mean	N	Std.
Format	1*	0.10	352	0.30
	2*	0.13	727	0.33
	3	0.15	768	0.36
	4*	0.19	718	0.40
	Total	0.15	2565	0.36
ANOVA	F=7.35	p=.00		

* The mean group difference is significant at the .05 level

Table 6. Duration interview in seconds, mean and standard deviations, Time 1

		Mean	N	Std.
Format	1*	455.75	347	390.91
	2	400.47	717	374.57
	3*	366.93	753	261.39
	4	392.29	700	499.49
	Total	395.78	2517	389.14
ANOVA	F=4.20	p=.01		

* The mean group difference is significant at the .05 level

Table 7. Overall evaluation by linear regression on evaluation questions

N=2565	Parameter	Std. error	t
Constant*	3.028	0.107	28.234
Duration*	0.198	0.013	14.784
Clarity*	0.145	0.018	7.826
Ease*	0.125	0.017	7.162
Layout*	0.104	0.011	9.119

*p<0.001

Table 8. How do you evaluate the duration of the interview? Mean and standard deviations at Time 2

		Mean	N	Std.
Format	1*	7.38	551	1.27
	2*	7.62	614	0.98
	3	7.55	644	1.14
	4*	7.45	529	1.12
	Total	7.51	2338	1.13
ANOVA	F=5.28	p=.00		

* The mean group difference is significant at the .05 level

Table 9. How easy were the questions to fill in? Mean and standard deviations at Time 2

		Mean	N	Std.
Format	1*	7.76	476	1.05
	2	7.63	537	0.98
	3*	7.57	559	1.14
	4*	7.57	442	0.99
	Total	7.63	2014	1.05
ANOVA	F=3.41	p=.02		

* The mean group difference is significant at the .05 level

Table 10. How was the layout? Mean and standard deviations at Time 1

		Mean	N	Std.
Format	1*	7.67	350	1.550
	2*	7.57	723	1.447
	3*	7.37	764	1.540
	4*	7.06	711	1.720
	Total	7.38	2548	1.584
ANOVA	F=17.13	p=.00		

* The mean group difference is significant at the .05 level

Table 11. Evaluation layout at Time 2 compared with Time 1

	Time 2		Time 1		
	Mean	Std.	N	Mean	Std.
Same number of items per screen at Time 2*	7.38	1.37	842	7.43	1.49
More items per screen at Time 2*	7.34	1.42	458	7.68	1.44
Fewer items per screen at Time 2*	7.59	1.08	715	7.21	1.76
Total	7.45	1.29	2015	7.41	1.59
ANOVA	F=7.28	p=.00		F=12.60	p=.00

* The mean group difference is significant at the .05 level

Table 12. Evaluation layout by linear regression on sex, education, age and interaction terms

N=2565	Parameter	Std. Error	t
Constant**	6.942	.412	16.848
Format2	-.244	.525	-.564
Format3	-.740	.505	-1.466
Format4*	-1.593	.512	-3.113
Sex	.102	.164	.619
Education	.015	.005	3.087
Age**	-.033	.054	-.606
Format2-Sex	.005	.201	.026
Format3-Sex	.038	.198	.190
Format4-Sex	.199	.201	.988
Format2-Educ	.032	.067	.484
Format3-Educ	.026	.065	.397
Format4-Educ	-.028	.066	-.430
Format2-Age	-.008	.006	-0.14
Format3-Age	.007	.006	1.120
Format4-Age*	.017	.006	2.825

*p<.01

**p<0.001

Note: for the definition of sex and education see Table 3.